**ANALISIS KOMPARASI PEMODELAN TOPIK TERJEMAH INDONESIA HADIST BUKHORI DENGAN BERTOPIC, LATEN SEMANTIC ANALYSIS (LSA) DAN LATENT DIRECHLET ALLOCATION (LDA)**

**Proposal Tesis**

**Diajukan untuk memenuhi syarat memperoleh gelar Magister Komputer pada Program Studi Teknik Informatika S-2**

****

OLEH :

**ASEP RIDWAN HIDAYAT**

**231012050036**

**PROGRAM STUDI TEKNIK INFORMATIKA S-2**

**PROGRAM PASCASARJANA**

**UNIVERSITAS PAMULANG**

**TANGERANG SELATAN**

**2024**

LEMBAR PERSETUJUAN PROPOSAL TESIS

**ANALISIS KOMPARASI PEMODELAN TOPIK TERJEMAH INDONESIA HADIST BUKHORI DENGAN BERTOPIC, LATEN SEMANTIC ANALYSIS (LSA) DAN LATENT DIRECHLET ALLOCATION (LDA)**

Telah disetujui untuk disidangkan pada Program Studi Teknik Informatika S-2

Program Pascasarjana Universitas Pamulang

Pada tanggal …………………

Oleh

Nama Asep Ridwan Hidayat

231012050036

Proposal Tesis ini telah disetujui oleh Pembimbing untuk ujian proposal:

|  |  |  |
| --- | --- | --- |
| Pembimbing I |  | Pembimbing II |
|  |  |  |
| Dr. Tukiyat, M.Si |  | Dr. Taswanda Taryo, M.Sc. |
| NIDN. 8898211019 |  | NIDN. |
|  |  |  |
| Disahkan:  Direktur Program Pascasarjana  Universitas Pamulang | | |
|  |  |  |
| Dr. Sajarwo Anggai, MT. | | |
| NIDN. xxxxxxx | | |

LEMBAR PENGESAHAN TESIS

**ANALISIS KOMPARASI PEMODELAN TOPIK TERJEMAH INDONESIA HADIST BUKHORI DENGAN BERTOPIC, LATEN SEMANTIC ANALYSIS (LSA) DAN LATENT DIRECHLET ALLOCATION (LDA)**

Telah dipertahankan di hadapan Dewan Penguji Program Program Studi Teknik Informatika S-2 Program Pascasarjana Universitas Pamulang

Pada tanggal …………………

Oleh  
Nama Mahasiswa

2210120000015

|  |  |  |
| --- | --- | --- |
| Penguji I |  | Penguji II |
| ……………………………………… |  | ……………………………………… |
| NIDN. ……………………………… |  | NIDN. ……………………………… |
|  |  |  |
| Pembimbing I |  | Pembimbing II |
| …………………………………… |  | ……………………………………… |
| NIDN. ……………………………… |  | NIDN. ……………………………… |
|  |  |  |
| Disahkan:  Direktur Program Pascasarjana  Universitas Pamulang | | |
|  |  |  |
| ……………………………………… | | |
| NIDN. ……………………………… | | |

LEMBAR PERNYATAAN TESIS

Dengan ini saya menyatakan bahwa dalam tesis ini tidak terdapat karya yang pernah diajukan untuk memperoleh gelar kesarjanaan di suatu Perguruan Tinggi, dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang lain, kecuali yang secara tertulis diacu dalam naskah tesis ini dan disebutkan dalam daftar pustaka.

Tangerang Selatan, September 2024

Tanda tangan

(Materai Rp. 10.000)

**Asep Ridwan Hidayat**

NIM : 231012050036

KATA PENGANTAR

Puji syukur penulis panjatkan ke hadirat Allah SWT, karena atas berkat dan rahmat -Nya penulis dapat menyelesaikan tesis ini. Penulis menyadari bahwa tanpa bantuan dan bimbingan dari berbagai pihak, mulai dari masa perkuliahan hingga penyusunan tesis ini, sangatlah sulit bagi penulis untuk menyelesaikannya. Oleh karena itu, penulis ingin mengucapkan terima kasih kepada:

1. Universitas Pamulang dan Program Studi Teknik Informatika S-2 yang telah melayani proses akademik dan pembelajaran dengan baik, mulai dari pendaftaran mahasiswa baru, pelaksanaan perkuliahan, hingga penyusunan tugas akhir.
2. Rektor Universitas Pamulang yang telah mengizinkan penulis untuk menempuh studi program S-2.
3. Dr. Sajarwo, Anggai, MT, selaku Ketua Program Studi Teknik Informatika S-2 Universitas Pamulang.
4. Dr. Tukiyat, M.Si. dan Dr. Taswanda Taryo, M.Sc.selaku dosen pembimbing I dan II yang telah menyediakan waktu, tenaga, dan pikiran untuk mengarahkan penulisan dan penyusunan tesis ini.
5. Rekan-rekan kerja serta mahasiswa Program Studi Teknik Informatika S-2 Universitas Pamulang yang telah banyak mendukung penulis dalam menyelesaikan tesis ini.
6. Semua pihak yang terlibat dan tidak penulis sebutkan satu per satu.

Akhir kata, penulis berharap Tuhan Yang Maha Esa berkenan membalas segala kebaikan semua pihak yang telah membantu. Semoga tesis ini membawa manfaat bagi pengembangan ilmu pengetahuan.

Penulis

Asep Ridwan Hidayat

PERNYATAAN PERSETUJUAN PUBLIKASI TESIS

Saya yang bertanda tangan di bawah ini:

Nama : Asep Ridwan Hidayat

NPM : 231012050036

Program Studi : Teknik Informatika S-2

Jenis Karya : Tesis

Sebagai sivitas akademik Universitas Pamulang, dengan ini saya menyetujui untuk memberikan kepada Universitas Pamulang Hak Bebas Royalti Noneksklusif (*Non-exclusive Royalty-Free Right*) atas karya ilmiah saya yang berjudul:

"ANALISIS KOMPARASI PEMODELAN TOPIK TERJEMAH INDONESIA HADIST BUKHORI DENGAN *BERTOPIC*, *LATEN SEMANTIC ANALYSIS* (LSA) DAN *LATENT DIRECHLET ALLOCATION* (LDA)

"

Beserta perangkat yang ada (jika diperlukan). Dengan Hak Bebas Royalti Noneksklusif ini, Universitas Pamulang berhak menyimpan, mengalih media/formatkan, mengelola dalam bentuk pangkalan data (*database*), merawat, dan mempublikasikan tesis saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan pemilik Hak Cipta.

Demikian pernyataan ini saya buat dengan sebenarnya.

Dibuat di: Tangerang Selatan

Pada tanggal: .....

Yang menyatakan,

(Asep Ridwan Hidayat)

ABSTRAK

Penelitian ini bertujuan untuk menganalisis dan membandingkan performa tiga metode pemodelan topik, yaitu *BerTopic*, *Latent Semantic Analysis* (LSA), dan *Latent Dirichlet Allocation* (LDA), dalam mengelompokkan dan mengidentifikasi topik-topik dari terjemahan Indonesia Hadis Bukhari. Pemodelan topik merupakan salah satu pendekatan penting dalam pemrosesan bahasa alami yang memungkinkan pemahaman lebih mendalam terhadap konten teks dengan mengekstraksi tema-tema utama. *BerTopic* adalah metode yang memanfaatkan representasi teks berbasis transformer dan clustering hierarkis untuk menghasilkan topik yang lebih akurat dan interpretatif. LSA, sebagai teknik berbasis dekomposisi matriks, mengidentifikasi hubungan laten antar kata dan dokumen dengan mengurangi dimensi data. Sementara itu, LDA merupakan model probabilistik yang memodelkan dokumen sebagai campuran dari beberapa topik dan mengasumsikan distribusi kata pada topik tertentu. Dalam penelitian ini, ketiga metode tersebut diterapkan pada dataset terjemahan Hadis Bukhari, dan hasilnya dibandingkan berdasarkan beberapa metrik seperti koherensi topik, akurasi klasifikasi, dan kemudahan interpretasi. Hasil analisis menunjukkan bahwa setiap metode memiliki kelebihan dan kekurangan yang berbeda, di mana BerTopic unggul dalam hal presisi topik, LSA memberikan dimensi laten yang abstrak, sementara LDA menghasilkan topik yang lebih terstruktur. Kesimpulan dari penelitian ini memberikan wawasan tentang metode terbaik untuk pemodelan topik teks religius dengan bahasa yang kompleks.

Kata Kunci: Topik Model, *BerTopic*, *Latent Semantic Analysis*, *Latent Dirichlet Allocation*, *Hadist Bukhari*

*ABSTRACT*

This study aims to analyze and compare the performance of three topic modeling methods, namely BerTopic, Latent Semantic Analysis (LSA), and Latent Dirichlet Allocation (LDA), in clustering and identifying topics from the Indonesian translation of Hadith Bukhari. Topic modeling is one of the important approaches in natural language processing that allows a deeper understanding of text content by extracting key themes. BerTopic is a method that utilizes transformer-based text representation and hierarchical clustering to generate more accurate and interpretative topics. LSA, as a matrix decomposition-based technique, identifies latent relationships between words and documents by reducing the dimensionality of the data. Meanwhile, LDA is a probabilistic model that models documents as a mixture of several topics and assumes the distribution of words on a particular topic. In this study, the three methods are applied to the Hadith Bukhari translation dataset, and the results are compared based on several metrics such as topic coherence, classification accuracy, and ease of interpretation. The analysis results show that each method has different advantages and disadvantages, where BerTopic excels in topic precision, LSA provides abstract latent dimensions, and LDA produces more structured topics. The conclusion of this study provides insight into the best method for topic modeling of religious texts with complex language.

*Keywords: Topic Model, BerTopic, Latent Semantic Analysis, Latent Dirichlet Allocation, Hadith Bukhari*

DAFTAR ISI

[LEMBAR PERSETUJUAN PROPOSAL TESIS ii](#_Toc177735747)

[LEMBAR PENGESAHAN TESIS iii](#_Toc177735748)

[LEMBAR PERNYATAAN TESIS iv](#_Toc177735749)

[KATA PENGANTAR v](#_Toc177735750)

[PERNYATAAN PERSETUJUAN PUBLIKASI TESIS vi](#_Toc177735751)

[ABSTRAK (200 kata – 300 kata) vii](#_Toc177735752)

[*ABSTRACT* viii](#_Toc177735753)

[DAFTAR ISI ix](#_Toc177735754)

[DAFTAR TABEL xii](#_Toc177735755)

[DAFTAR GAMBAR xiv](#_Toc177735756)

[BAB I PENDAHULUAN 1](#_Toc177735757)

[1.1 Latar Belakang 1](#_Toc177735758)

[1.2 Permasalahan Penelitian 7](#_Toc177735759)

[1.2.1 Identifikasi Masalah 7](#_Toc177735760)

[1.2.2 Ruang Lingkup Masalah 8](#_Toc177735761)

[1.2.3 Rumusan Masalah 8](#_Toc177735762)

[1.3 Tujuan dan Manfaat Penelitian 8](#_Toc177735763)

[1.3.1 Tujuan Penelitian 8](#_Toc177735764)

[1.3.2 Manfaat Penelitian 9](#_Toc177735765)

[1.4 Sistematika Penulisan 10](#_Toc177735766)

[BAB II LANDASAN TEORI 12](#_Toc177735767)

[2.1 Tinjauan Pustaka 12](#_Toc177735768)

[((konsep yang ada ditulis SLR dan ada tabel). Sitasi minimal 30 – 40 paper, buku dam sumber ilmiah yang dapat dipercaya) 12](#_Toc177735769)

[2.2 Landasan Teori 23](#_Toc177735770)

[2.2.1 Analisis Sentimen 23](#_Toc177735771)

[2.2.2 Jasa Pengiriman 24](#_Toc177735772)

[2.2.3 JNE 25](#_Toc177735773)

[2.2.4 Alur Pengiriman JNE 27](#_Toc177735774)

[2.2.5 *Google Play* 29](#_Toc177735775)

[2.2.6 *Online Review* 30](#_Toc177735776)

[2.2.7 *Web Scraping* 31](#_Toc177735777)

[2.2.8 *Machine Learning* 31](#_Toc177735778)

[2.2.9 *Text Mining* 32](#_Toc177735779)

[2.2.10 *Text Preprocessing* 33](#_Toc177735780)

[2.2.11 *Clasification* 33](#_Toc177735781)

[2.2.12 *Data Mining* 35](#_Toc177735782)

[2.2.13 *Naïve Bayes Clasifier* 35](#_Toc177735783)

[2.2.14 *KNN* 37](#_Toc177735784)

[2.2.15 Fitur dan Pembobotan 38](#_Toc177735785)

[2.2.16 Asosiasi Kata 40](#_Toc177735786)

[2.2.17 *Diagram Fishbone* 40](#_Toc177735787)

[2.2.18 *Google Colab* 41](#_Toc177735788)

[2.3 Kerangka Pemikiran 43](#_Toc177735789)

[2.4 Hipotesis 44](#_Toc177735790)

[BAB III METODE PENELITIAN 45](#_Toc177735791)

[3.1 Analisis Kebutuhan 45](#_Toc177735792)

[3.1.1 Populasi dan Sampel 45](#_Toc177735793)

[3.1.2 Metode Pengumpulan Data 45](#_Toc177735794)

[3.1.3 Jenis dan Sumber Data 46](#_Toc177735795)

[3.1.4 Perancangan Penelitian 47](#_Toc177735796)

[3.1.5 Metode Analisis Data 49](#_Toc177735797)

[3.2 *Scraping Data* 49](#_Toc177735798)

[3.3 *Preprocessing* 50](#_Toc177735799)

[3.3.1 Model Akurasi Data *dengan Confusion Matrix* 50](#_Toc177735800)

[3.3.2 Akurasi (*Accuracy*) 51](#_Toc177735801)

[3.3.3 Presisi (*Precision*) 51](#_Toc177735802)

[3.3.4 *Recall* (Sensitivitas atau *True* *Positive* *Rate*) 51](#_Toc177735803)

[3.3.5 *F1 Score* 52](#_Toc177735804)

[DAFTAR PUSTAKA 53](#_Toc177735805)

DAFTAR TABEL

[**Tabel 1.1.** Data *Top Brand* Jasa Kurir Indonesia 3](#_Toc177735806)

DAFTAR GAMBAR

[Gambar 2.1. Tahapan Proses Pengiriman JNE 27](#_Toc177736102)

[Gambar 2.2. Kerangka Pemikiran 43](#_Toc177736103)

1. PENDAHULUAN
   1. Latar Belakang

Hadis merupakan salah satu sumber utama ajaran Islam setelah Al-Qur'an yang memberikan pedoman hidup bagi umat muslim. Dalam konteks perkembangan teknologi informasi, pengolahan teks hadis secara digital semakin dibutuhkan untuk mendukung penelitian dan aplikasi keagamaan. Salah satu tantangan dalam analisis teks hadis, khususnya terjemahan hadis, adalah memahami struktur topik yang terkandung di dalamnya. model topik menawarkan pendekatan yang berbeda dalam melakukan pengelompokan topik dari kumpulan teks.

Pemodelan topik adalah teknik analisis teks yang digunakan untuk menemukan topik tersembunyi dalam kumpulan dokumen. Topik dapat didefinisikan sebagai sekelompok kata yang sering muncul bersama dan mewakili ide atau konsep tertentu. Prinsip dasar pemodelan topik adalah bahwa setiap dokumen dapat dianggap sebagai campuran dari beberapa topik. Setiap topik memiliki sekumpulan kata yang terkait dengannya, dan setiap kata dalam dokumen dapat dikaitkan dengan salah satu topik tersebut. (Blei, 2012)

Metode awal dalam pemodelan topik menggunakan pendekatan probabilistik (Blei et al., 2003), sedangkan metode modern saat ini telah memanfaatkan teknik pembelajaran mendalam untuk meningkatkan pemahaman model terhadap teks yang digunakan (Churchill & Singh, 2022). BERTopic merupakan sebuah pustaka pemodelan topik modern yang menggunakan model bahasa SBERT (Reimers & Gurevych, 2019) sehingga pemahaman model terhadap konteks dari kata-kata meningkat. Di dalam artikelnya BERTopic membandingkan kinerja algoritmanya dengan LDA berdasarkan nilai koherensi dan nilai keberagaman. Hasil dari evaluasi tersebut adalah BERTopic memiliki nilai yang lebih besar dibandingkan LDA. (Grootendorst, 2022)

Seperti *Latent Dirichlet Allocation* (LDA) dan BERTopic menawarkan pendekatan yang berbeda dalam melakukan pengelompokan topik dari kumpulan teks. LDA menggunakan pendekatan statistik berdasarkan distribusi kata, sementara BERTopic menggabungkan model embedding berbasis BERT (*Bidirectional* *Encoder Representations from Transformers*) yang lebih kontekstual. Oleh karena itu, komparasi kedua metode ini pada terjemahan hadis Bukhari dalam bahasa Indonesia menjadi relevan untuk diinvestigasi guna menemukan pendekatan yang paling tepat dalam mengekstraksi dan mengelompokkan topik-topik dari teks hadis.

Analisis komparasi pemodelan topik *latent dirichlet allocation* (LDA) dan *latent semantic analysis* (LSA) sudah dilakukan oleh beberapa orang peneliti. Yaitu penelitian Ulfa Mulya yang melakukan Analisis komparasi pemodelan topik *latent dirichlet allocation* dan *latent semantic analysis* pada ulasan restoran di Yogyakarta. Kesimpulan penelitian ini adalah *Latent Dirichlet Allocation* (LDA) *dan Latent Semantic Analysis* (LSA) sama-sama menghasilkan model topik yang baik, namun luaran model terbaik dihasilkan oleh LDA. Dalam studi kasus ini, LDA terbukti bekerja lebih baik dari LSA. Dengan tujuh topik menjadi topik luaran terbaik, yaitu topik 1 membahas tentang “harga”, topik 2 membahas tentang “suasana”, topik 3 membahas tentang “menu”, topik 4 membahas tentang “keunikan”, topik 5 membahas tentang “lesehan”, topik 6 membahas tentang “instagramable”, dan topik 7 membahas tentang “nongkrong”.

Penelitian selanjutnya yang mirip dilakukan oleh Herwinsyah penelitian ini melakukan pemodelan topik dalam al-qur'an menggunakan library bertopic pada model bahasa Bert. Metode pemodelan topik menggunakan BERTOPIC. Hasil penelitian adalah pemodelan topik menghasilkan 8 topik utama secara terperinci sebagai berikut; Topic 0 Al Quran dengan prosentase sebesar 6%, Topic 1 Aku (Allah) sebesar 6,5%, Topic 2 Langit 3,8%, Topic 3 Rasul 8%, Topic 4 Malaikat 12,5%, Topic 5 Wanita 5%, Topic 6 Neraka dengan prosentase 13%, serta Topic 7 Dibangkitkan sebesar 5,5%. Kata-kata tersebut dianggap sangat penting dalam mewakili topik-topik yang dihasilkan dan juga termasuk dalam kategori spiritual, moral, dan hukum.

Perbedaan penelitian yang sudah dilakukan sebelumnya dengan penelitian ini adalah pada penelitian ini dilakukan dengan tiga metode topik model yaitu Bert, *latent dirichlet allocation* (LDA) dan *latent semantic analysis* (LSA). Pengambilan data juga dilakukan dengan cara mengambil *Dataset*.

Pada penelitian kali ini, penulis akan melakukan komparasi algoritma topic modeling, yaitu Latent Dirichlet Allocation, Latent Semantic Analysis dan Bertopic. Hasil komparasi diharapkan dapat ditemukannya algoritma terbaik dengan nilai koherensi yang tinggi sehingga menghasilkan luaran topik model yang produktif, efektif dan memiliki makna jelas.

* 1. Permasalahan Penelitian

Berdasarkan latar belakang masalah diperoleh rumusan masalah penelitian yang akan diteliti sesuai dengan identifikasi dan dibatasi dalam suatu ruang lingkup permasalahan.

* + 1. Identifikasi Masalah

Berdasarkan latar belakang yang telah diuraikan, rumusan masalah yang akan diselesaikan dalam penelitian ini adalah:

1. Bagaimana kinerja BERTopic LDA dan LSA berdasarkan nilai koherensi, nilai keberagaman, dan waktunya?
   * 1. Ruang Lingkup Masalah

Berdasarkan pada latar belakang masalah yang ada maka penelitian ini akan membatasi permasalahan yang ada, yaitu :

1. Pemodelan topik LDA dilakukan menggunakan pustaka Gensim
2. Pemodelan topik BERTopic dilakukan menggunakan pustaka BERTopic
3. Korpus yang akan digunakan dataset terjemah Hadis Bukhori
4. Penelitian dilakukan menggunakan bahasa pemrograman Python versi 3.10
   * 1. Rumusan Masalah

Berdasarkan identifikasi masalah di atas, maka dapat dibuat rumusan masalah sebagai berikut :

1. Bagaimana kinerja model BerTopic, LSA, dan LDA dalam mengklasifikasikan topik pada teks terjemahan Hadis Bukhari dalam bahasa Indonesia?
2. Apa perbedaan hasil klasifikasi topik yang dihasilkan oleh BerTopic dibandingkan dengan LSA dan LDA dalam konteks teks terjemahan Hadis?
   1. Tujuan dan Manfaat Penelitian
      1. Tujuan Penelitian

Tujuan penelitian ini adalah untuk memberikan jawaban atas identifikasi masalah yang telah disebutkan diatas, yaitu :

1. Menganalisis dan membandingkan kinerja model BerTopic, Latent Semantic Analysis (LSA), dan Latent Dirichlet Allocation (LDA) dalam mengklasifikasikan topik pada teks terjemahan Hadith Bukhari berbahasa Indonesia.
2. Mengidentifikasi perbedaan hasil klasifikasi topik yang dihasilkan oleh BerTopic, LSA, dan LDA dalam konteks semantik dan makna mendalam dari Hadith Bukhari.
3. Memberikan rekomendasi model pemodelan topik yang terbaik dan paling sesuai untuk diterapkan pada teks-teks keagamaan, khususnya terjemahan Hadith Bukhari dalam bahasa Indonesia.
   * 1. Manfaat Penelitian
4. Manfaat bagi Penulis:

* Pengembangan Keahlian:

Penelitian ini memberikan kesempatan bagi penulis untuk mengembangkan keterampilan analisis data dan pemahaman mendalam mengenai metode topik model dengan BERtopic, *Laten Semantic Analysis* (LSA) dan *Latent Direchlet Allocation* (LDA).

* Peningkatan Kompetensi Akademik:

Melalui penelitian ini, penulis dapat memperkuat kompetensi akademik dan kemampuan dalam menyusun karya ilmiah yang berkualitas, yang akan berguna dalam karir akademik dan profesional di masa depan.

* Kontribusi terhadap Pengetahuan:

1. Penelitian ini memberikan kontribusi dalam memahami pola tematik dari Hadith Bukhari dalam bahasa Indonesia.
2. penelitian ini dapat menjadi rujukan bagi pengembangan sistem klasifikasi hadith berbasis NLP untuk bahasa lain, sehingga memperluas cakupan penerapan teknologi NLP dalam studi agama.
3. Manfaat bagi Kampus:

* Peningkatan Reputasi Akademik:

Penelitian ini akan berkontribusi pada peningkatan reputasi kampus dalam bidang riset, khususnya di bidang teknologi informasi dan analisis data, melalui publikasi hasil penelitian yang berkualitas.

* Sumber Referensi bagi Mahasiswa:

Hasil penelitian ini dapat menjadi sumber referensi bagi mahasiswa lain yang tertarik untuk melakukan penelitian dalam bidang yang sama, sehingga memperkaya literatur dan penelitian yang ada di kampus.

* 1. Sistematika Penulisan

Penulisan tesis ini disusun secara sistematis untuk memberikan gambaran yang jelas mengenai seluruh tahapan penelitian yang dilakukan. Sistematika penulisan ini terdiri dari lima bab utama, yang masing-masing diuraikan sebagai berikut:

**BAB I - Pendahuluan**

Bab ini berisi latar belakang penelitian, perumusan masalah, tujuan dan manfaat penelitian, serta sistematika penulisan. Pendahuluan memberikan konteks yang jelas mengenai pentingnya penelitian ini dan masalah-masalah yang ingin dipecahkan.

**BAB II - Landasan Teori dan Kerangka Pemikiran**

Berisi tentang Teori yang digunakan terdiri dari text mining, topic modeling, metode Latent Dirichlet Allocation dan Latent Semantic Analysis, text preprocessing, Bertopik, N Gram, Dictionary, Corpus dan Python.

**BAB III - Metode Penelitian**

Bab ini dipaparkan tentang metode yang peneliti pakai pengumpulan data yang akan dilakukan pada penelitian ini.

**BAB IV - Hasil dan Pembahasan**

Memaparkan hasil percobaan sesuai dengan metodologi yang dirancang pada bab sebelumnya.

**BAB V - Kesimpulan dan Saran**

Bab ini menyimpulkan hasil penelitian dan menjawab pertanyaan penelitian yang diajukan. Kesimpulan yang diambil berdasarkan temuan penelitian disampaikan secara ringkas dan jelas. Selain itu, bab ini juga memberikan saran untuk penelitian lanjutan serta rekomendasi praktis bagi pihak-pihak yang berkepentingan.

1. LANDASAN TEORI
   1. Tinjauan Pustaka

((konsep yang ada ditulis SLR dan ada tabel). Sitasi minimal 30 – 40 paper, buku dam sumber ilmiah yang dapat dipercaya)

Penelitian mengenai analisis sentimen telah berkembang pesat, terutama dalam konteks penggunaan algoritma machine learning seperti Naïve Bayes, K-Nearest Neighbors (KNN), dan Support Vector Machines (SVM). Berbagai studi telah dilakukan untuk menganalisis sentimen dari berbagai data teks, termasuk ulasan pengguna di platform media sosial, ulasan aplikasi, dan opini publik di media online. Metode dan hasil dari penelitian ini menunjukkan bagaimana berbagai algoritma digunakan untuk mengklasifikasikan data menjadi sentimen positif, negatif, atau netral, serta memberikan saran untuk penelitian lebih lanjut di bidang ini. Penelitian sebelumnya yang dijadikan referensi mencakup berbagai metode dan pendekatan dalam analisis sentimen, khususnya yang menggunakan data dari media sosial seperti *Twitter*. Penelitian-penelitian tersebut memberikan dasar teori dan metodologi yang kuat untuk penelitian ini, serta menginformasikan tentang teknik-teknik *Preprocesing* data, implementasi algoritma *Machine Learning*, dan evaluasi kinerja *Classifiers*.

Penelitian tentang *"Analisis Sentimen Pengguna Twitter terhadap Dompet Elektronik dengan Metode Lexicon Based dan K-Nearest Neighbors"* mengkaji sentimen pengguna terhadap layanan dompet elektronik seperti OVO, Gopay, dan LinkAja. Hasil dari penelitian ini menunjukkan bahwa metode K-Nearest Neighbors menghasilkan akurasi tertinggi pada Gopay sebesar 94,05%, sementara OVO dan LinkAja memiliki akurasi yang lebih rendah. Kesimpulan dari penelitian ini adalah metode K-Nearest Neighbors cukup efektif dalam mengklasifikasikan sentimen pengguna terhadap layanan dompet elektronik. Peneliti menyarankan agar penelitian lebih lanjut dilakukan dengan memperluas dataset untuk meningkatkan akurasi dan validitas model (Salim, 2020).

**Ricardo Cuatanto dan Rudi Sutomo (2020)** dalam penelitian yang berjudul *"Comparative Analysis Between Naïve Bayes Algorithm and Decision Tree Loss Rate from Fire Disaster Data in DKI Jakarta Province"* membandingkan kinerja dua algoritma, yaitu Naïve Bayes dan Decision Tree, dalam mengklasifikasikan tingkat keparahan kerugian akibat kebakaran di Jakarta. Penelitian ini menemukan bahwa algoritma Decision Tree lebih unggul dengan akurasi sebesar 78%, dibandingkan dengan Naïve Bayes yang hanya mencapai 75%. Hasil ini menunjukkan bahwa Decision Tree lebih efektif dalam mengelompokkan data terkait bencana kebakaran. Saran dari penelitian ini adalah untuk meningkatkan kesiapsiagaan dan mitigasi kebakaran di kawasan perkotaan dengan memanfaatkan hasil klasifikasi ini (Cuatanto & Sutomo, 2020).

Dan seterusanya dibuat minimal 10 peneliti terdahulu yang dijadikan sebagai acuan utama dalam penelitian ini,,,,,,

Lanjutkan ,,,,,,

Dari beberapa penelitian terdahulu yang dijadikan acuan pustaka, berikut ini diberikan reviu paper seperti pada tabel tabel 2.1 yang menyajikan ringkasan dari beberapa penelitian relevan atau penelitian sebelumnya.

**Tabel 2.1.** Tinjauan Pustaka

| **No** | **Penulis** | **Tahun** | **Judul Penelitian** | **Model** | ***Accuracy*** |
| --- | --- | --- | --- | --- | --- |
| 1 | Ss. Salim | 2020 | Analisis Sentimen pengguna *Twitter* terhadap Dompet eletronik dengan Metode *Lexicon Based* dan *K-Nearest Neighbors* | *Lexicon Based dan K-Nearest Neighbors* | Hasil akhir perhitungan metode *K-Nearest Neighbors* dengan *Confusion Matrix* untuk *tweet* OVO diperoleh nilai akurasi positif sebesar 86,91% dan nilai akurasi negatif 13,09%, *tweet* Gopay diperoleh nilai akurasi sebesar 94,05% dan nilai akurasi negatif 5,95%, serta *tweet* LinkAja diperoleh nilai akurasi sebesar 76,31% dan nilai akurasi *negative* 23,69%. |
| 2 | Ricardo Cuatanto, Rudi Sutomo | 2020 | *Comparative Analysis Between Naïve Bayes*  *Algorithm and Decision Tree Loss Rate from Fire Disaster Data in DKI Jakarta Province* | *Naïve Bayes* | penelitian ini membantu meningkatkan teknik mitigasi kebakaran. Keamanan dan kesiapan menghadapi bencana kebakaran di kawasan ini harus ditingkatkan. Akurasi kedua prediksi yang dilakukan algoritma *Naïve Bayes* adalah 75%. Sebaliknya, akurasi algoritma *Decision Tree* sebesar 78% sehingga menghasilkan kesimpulan bahwa pendekatan *Decision Tree* lebih membantu dalam mengkategorikan tingkat keparahan kerugian akibat bencana kebakaran. Kata Kunci Bencana Kebakaran Algoritma *Naïve Bayes* Algoritma *Decision Tree* Klasifikasi Tingkat Kerugian. |
| 3 | Debby Alita, Auliya Rahman Isnain | 2020 | Pendeteksian Sarkasme Pada Proses Analisis Sentimen Menggunakan *Random Forest* *Classifier* | *Random Forest* *Classifier* | Hasil penelitian ini didapatkan peningkatan nilai rata- rata akurasi sebesar 16,61 %, nilai presisi sebesar 5,45 %, nilai *Recall* sebesar 9,64% dan kenaikan nilai *F1-Score* sebesar 11,27% dengan jumlah data sebanyak 2.027 dengan rincian data dengan label positif berjumlah 1023, data dengan label negatif berjumlah 587 dan data dengan label netral berjumlah 462. Data sarkasme didapatkan dari *tweet* dengan label positif yang kemudian diberikan label sarkasme atau tidak sarkasme dan didapat hasil label dengan jumlah keseluruhan berlabel sarkasme |
| 4 | Harun & Putri Ananda | 2021 | Analisa Sentimen Opini Publik Tentang Vaksinasi Covid-19 di Indonesia Menggunakan *Naïve Bayes* dan *Decision Tree* | *Naïve Bayes* | analisa sentimen opini masyarakat tentang vaksinasi COVID-19 yang telah dilakukan, cenderung ke tanggapan negatif dengan nilai akurasi 100.00% menggunakan algoritma NBC dan 50.39% menggunakan algoritma *Naïve Bayes*. |
| 5 | RD Septiana, AB Susanto, T Tukiyat | 2021 | Analisis Sentimen Vaksinasi Covid-19 Pada *Twitter* Menggunakan *Naïve Bayes* *Classifier* Dengan *Feature Selection Chi- Squared Statistic* Dan *Particle Swarm Optimization* | *Chi-Squared Statistic, NBC, PSO* | Hasil Pengujian algoritma *Naïve Bayes* *Classifier* menggunakan seleksi fitur *Chi Squared Statistic* berbasis *forward selection* memiliki nilai akurasi yang lebih tinggi dibandingkan menggunakan seleksi. |
| 6 | H Setiawan, E Utami, S Sudarmawan | 2021 | Analisis Sentimen *Twitter* Kuliah *Online* Pasca Covid-19 Menggunakan Algoritma *Support Vector Machine* dan *Naïve Bayes* | *Naïve Bayes, KNN* | Untuk analisis sentimen peneliti menerapkan algorima naïve *Naïve Bayes* dan *KNN* dengan hasil peforma yang didapat pada algoritma *Naïve Bayes* akurasi 81,20%, waktu 9,00 detik, *Recall* 79,60% dan presisi 79,40% sedangkan untuk algoritma *KNN* mendapatkan nilai akurasi 85%, waktu 31,60 detik, *Recall* 84% dan presisi 83,60%, hasil peforma tersebut diperoleh pada iterasi ke 1 untuk naïve *Naïve Bayes* dan iterasi ke 423 untuk algoritma *KNN*. |
| 7 | S Samsir, A Ambiyar, U Verawardina, F Edi, R Watrianthos | 2021 | Analisis Sentimen Pembelajaran Daring Pada *Twitter* di Masa Pandemi COVID-19 Menggunakan Metode *Naïve Bayes* | *Naïve Bayes* | Menunjukkan bahwa pembelajaran daringmemiliki 30% sentimen positif, 69% sentimen negatif, dan 1% netral pada periode tersebut. Tingginya sentimen negatif dihasilkan karena ketidakpuasaan masyarakat terhadap pembelajaran daring. Beberapa twit menunjukkan kekecewaan dengan kata „stres‟ dan „malas‟ merupakan kata yang memiliki frekuensi tinggi dalam percakapan. |
| 8 | Wijaya & Dwiasnati | 2020 | Implementasi *Data Mining* dengan Algoritma *Naïve Bayes* pada Penjualan Obat | *Naïve Bayes* | Penelitian ini menggunakan *tools* *Rapidminner* versi 8 sebagai media untuk mengujidata yang akan diolah untuk mendapatkan hasil *Accuracy* dan nilai *ROC*. Nilai *Accuracy* tersebut menunjukkan di nilai 88.00%. |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

* 1. Landasan Teori
     1. Analisis Sentimen

Analisis Sentimen atau opinion mining merupakan studi komputasi mengenai pendapat, perilaku, dan emosi seseorang terhadap entitas yang mana entitas tersebut dapat mewakili individu, kejadian atau topik (Indrayuni, 2018). Melalui analisis sentimen dapat diketahui opini publik terkait suatu hal baik berupa produk, jasa, aplikasi, film, musik, dan lain sebagainya melalui *review* pengguna pada berbagai situs. Analisis sentimen digunakan untuk melihat kecenderungan apakah pendapat atau opini pada suatu masalah cenderung *term*asuk dalam opini positif atau opini negatif (Kurniawan & Susanto, 2019).

Dlam konteks ini tuliskan teori-teori yang akan digunakan dan mendukung untuk penelitian saudara

Misal Teori Data *Mining*

Data mining adalah proses untuk mengekstrak informasi atau pengetahuan yang berguna dari kumpulan data yang besar. Secara umum, *Data Mining* melibatkan penerapan teknik-teknik statistik, matematika, dan *Machine Learning* untuk mengidentifikasi pola tersembunyi, tren, atau hubungan dalam data yang dapat digunakan untuk membuat keputusan yang lebih baik. Data mining berfungsi sebagai alat penting dalam berbagai bidang, *term*asuk bisnis, keuangan, pemasaran, kesehatan, dan penelitian ilmiah, untuk menganalisis data yang besar dan kompleks serta menemukan wawasan yang sebelumnya tidak terlihat.

Proses *Data Mining* terdiri dari beberapa tahap, mulai dari pengumpulan data, pembersihan dan persiapan data, hingga penerapan algoritma untuk menemukan pola atau model dalam data. Teknik-teknik yang digunakan dalam *Data Mining* antara lain *clustering*, *class*ification, *association rule mining*, dan *anomaly detection*. *Clustering* adalah proses pengelompokan data berdasarkan kesamaan, sedangkan *classification* melibatkan penentuan label atau kategori untuk data baru berdasarkan model yang sudah dilatih. *Association rule mining* digunakan untuk menemukan hubungan antara variabel dalam *Dataset*, dan *anomaly detection* bertujuan untuk mengidentifikasi data yang menyimpang dari pola umum.

Data mining menjadi semakin relevan seiring dengan pertumbuhan data yang cepat dalam era digital saat ini. Dengan kemampuannya untuk mengolah dan menganalisis data dalam jumlah besar, *Data Mining* memungkinkan organisasi dan individu untuk membuat keputusan berbasis data yang lebih informatif dan akurat. Di bidang penelitian, *Data Mining* digunakan untuk menemukan pola atau hubungan dalam data yang dapat memberikan wawasan baru atau mengonfirmasi hipotesis yang ada. Dengan demikian, *Data Mining* tidak hanya membantu dalam pengambilan keputusan, tetapi juga mendorong inovasi dan penemuan baru dalam berbagai disiplin ilmu.

* + 1. *Naïve Bayes Clasifier*

*Naïve Bayes* *Classifier* merupakan teknik klasifikasi berdasarkan Teorema *Bayes* dengan asumsi independensi di antara para prediktor. *Naïve Bayes* *Classifier* memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya sehingga dikenal sebagai Teorema *Bayes*. *Naïve Bayes* merupakan metode paling sederhana dari pengklasifikasian probabilitas, memiliki tingkat akurasi yang sangat tinggi ketika diaplikasikan pada database dengan big data (Wati, 2016).Menurut Rennie et al.(2003) Menyatakan bahwa *Naïve Bayes* adalah salah satu model yang paling efisien dan tangguh dalam kalasifikasi text.yang di tekankan bahwa meskipun asumsi independensi yang mendasari model ini jarang terpenuhi,perporma yang dihasilkan sering memuaskan dalam aplikasi praktis.

Selain itu model naïve *Naïve Bayes* juga memiliki tingkat kesalahan yang sangat minimum dibandingkan dengan algoritma klasifikasi lainnya (Liu etal., 2016).

Algoritma naïve *Naïve Bayes* hanya membutuhkan jumlah data pelatihan (*training data*) yang kecil untuk menentukan estimasi parameter yang diperlukan dalam proses pengklasifikasian. Karena yang diasumsikan sebagai variabel independent, maka hanya variansi dari suatu variabel dalam sebuah kelas yang dibutuhkan untuk menentukan klasifikasi, bukan keseluruhan dari matriks kovarians. Rumus *Bayes* dalam (Muslehatindkk, 2017) secara umum adalah sebagai berikut:

(1)

dengan :

X = Data dengan *class* yang belum diketahui

H = Hipotesis data X merupakan suatu *class* spesifik

P(H|X) = Probabilitas hipotesis H berdasarkan kondisi x (posteriori prob.)

P(H) = Probabilitas hipotesis H (*prior prob*) P(X|H) = Probabilitas X berdasarkan kondisi tersebut

P(X) = Probabilitas dari X

Adapun aturan *Bayes* adalah sebagai berikut :

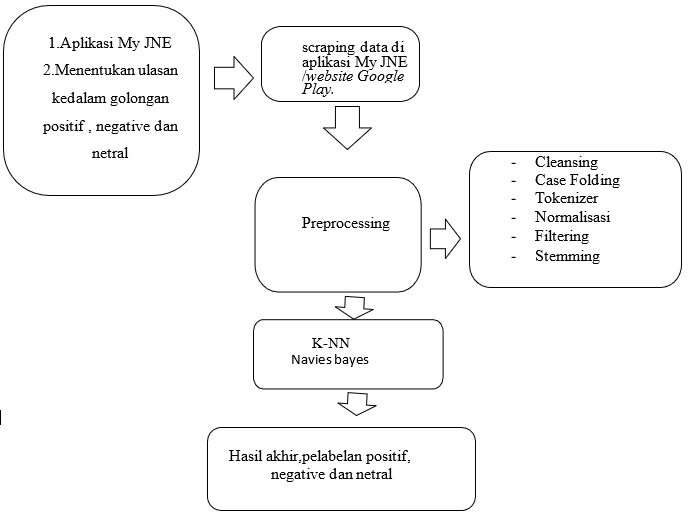
Jika P(h1|x) < P(h2|x), maka x diklasifikasikan sebagai h2. Pernyataan P(h1|x) mengindikasikan probabilitas hipotesis h1 berdasarkan kondisi x terjadi, begitu pula dengan h2. Sehingga didapat klasifikasi dari x sesuai dengan probabilitas terbesar diantara probabilitas x terhadap semua kelas.

Algoritma *Naïve Bayes* mengasumsikan bahwa fitur-fitur dalam data bersifat independen secara kondisional terhadap kelas yang diberikan. Artinya, setiap fitur dianggap tidak mempengaruhi fitur lainnya setelah mempertimbangkan kelas.Untuk mengklasifikasikan data baru, algoritma menghitung probabilitas posterior untuk setiap kelas menggunakan teorema *Bayes* dan memilih kelas dengan probabilitas tertinggi.

Ada beberapa Jenis *Naïve Bayes*, Gaussian *Naïve Bayes* Digunakan ketika fitur-fitur dalam data diasumsikan mengikuti distribusi Gaussian (normal). Biasanya digunakan untuk data yang bersifat kontinu.Multinomial *Naïve Bayes* Cocok untuk data yang berupa count atau frekuensi, seperti dalam klasifikasi teks. Misalnya, menghitung berapa kali suatu kata muncul dalam dokumen.Bernoulli *Naïve Bayes* Digunakan ketika fitur adalah variabel biner (misalnya, hadir/tidak hadir). Sering digunakan dalam klasifikasi teks dengan representasi biner.

* 1. Kerangka Pemikiran

Berikut ini adalah kerangka pemikiran berdasarkan hasil identifikasi masalah dan dilakukan tahapan penyesuaian berdasarkan dasar teori yang telah diidentifikasi sebelumnya. Kerangka pemikiran penelitian sebagai berikut.



**Gambar 2.2.** Kerangka Pemikiran

Diagram alur seperti pada gambar 2.2. menggambarkan proses pengolahan data dan analisis sentimen dimulai dari aplikasi My JNE, yang berfungsi sebagai titik awal pengumpulan data. Data diperoleh melalui teknik scraping, yang mengacu pada ekstraksi informasi dari aplikasi My JNE dan situs web Google Play. Proses pengumpulan data ini memastikan bahwa informasi yang diperoleh adalah representatif dan relevan untuk analisis lebih lanjut.

Setelah data terkumpul, tahap preprocessing dimulai untuk mempersiapkan data agar siap dianalisis. Proses pertama dalam tahap ini adalah cleansing, di mana data dibersihkan dari karakter atau simbol yang tidak diperlukan, sehingga hanya teks yang bersih dan relevan yang dipertahankan. Selanjutnya, case folding diterapkan untuk mengubah seluruh teks menjadi huruf kecil, memastikan konsistensi dalam analisis dan menghindari perbedaan antara huruf kapital dan huruf kecil. Proses ini diikuti oleh tokenization, yang memecah teks menjadi unit-unit kecil seperti kata atau frasa, yang memungkinkan analisis yang lebih mendalam pada level kata.

Tahap berikutnya adalah normalisasi, yang mengubah kata-kata menjadi bentuk dasar atau standar mereka, mengurangi variasi dalam data teks. Filtering dilakukan untuk menghapus kata-kata yang tidak relevan atau terlalu umum, seperti stop words, yang tidak memberikan informasi tambahan dalam konteks analisis sentimen. Terakhir, stemming diterapkan untuk mengubah kata-kata menjadi bentuk dasar atau akar katanya, menyederhanakan teks dan memudahkan proses analisis lebih lanjut.

Setelah preprocessing, data siap untuk dianalisis menggunakan metode analisis yang telah dipilih. Di antara metode yang digunakan adalah K-Nearest Neighbors (K-NN), yang merupakan algoritma klasifikasi berbasis kedekatan data, dan Naive Bayes, yang merupakan algoritma probabilistik yang mengklasifikasikan data berdasarkan distribusi probabilitas. Metode ini memanfaatkan fitur-fitur yang telah diproses untuk menentukan kategori sentimen dari data.

Hasil akhir dari proses ini adalah pelabelan sentimen, di mana setiap data dikategorikan sebagai positif, negatif, atau netral berdasarkan analisis yang telah dilakukan. Diagram ini menggambarkan alur kerja umum dalam pemrosesan bahasa alami (NLP) untuk analisis sentimen, di mana data mentah melalui beberapa tahap pemrosesan dan analisis untuk menghasilkan informasi yang dapat digunakan untuk memahami sentimen dari data yang dianalisis.

* 1. Hipotesis

(jika ada)

1. METODE PENELITIAN
   1. Analisis Kebutuhan

Secara lebih spesifik, objek penelitian mencakup:

1. Korpus teks terjemahan Hadith Bukhari: Hadith-hadith dalam bahasa Indonesia yang akan dianalisis.
2. Metode pemodelan topik: BerTopic, LSA, dan LDA, yang diterapkan pada korpus teks untuk menemukan dan mengelompokkan topik-topik dalam hadith.

Objek ini dipilih untuk mengevaluasi efektivitas ketiga metode dalam klasifikasi topik yang dapat membantu dalam pemahaman dan pengelolaan teks Hadith Bukhari.

* + 1. Metode Pengumpulan Data

Metode yang digunakan untuk proses pengumpulan data dari penelitian ini adalah sebagai berikut :

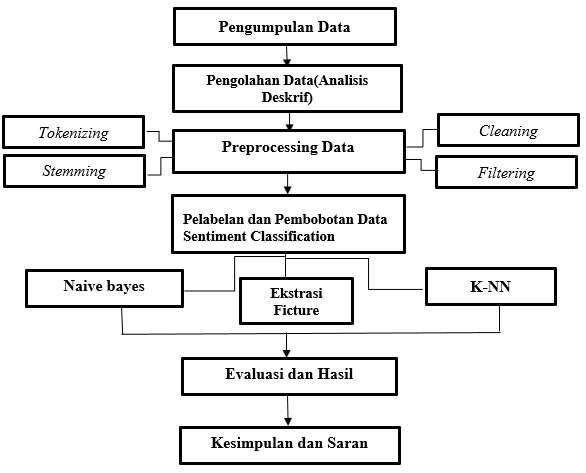
1. Data dikumpulkan dari terjemahan Hadith Bukhari yang tersedia secara digital.
2. Selanjutnya, teks akan diproses melalui beberapa tahap berikut: Preprocessing Teks: Teks akan dibersihkan dari karakter khusus, kata-kata yang tidak relevan, serta dilakukan tokenisasi dan lemmatization untuk mendapatkan bentuk kata dasar. Stopwords dalam bahasa Indonesia juga akan dihapus.
3. Pembagian Teks Hadith Bukhari akan dibagi menjadi dokumen-dokumen individual (berdasarkan bab atau topik tertentu dalam hadith) yang siap untuk dianalisis menggunakan model pemodelan topik.Studi Pustaka.
4. Studi pustaka dilakukan dengan menggunakan beberapa kajian literatur, buku, maupun referensi jurnal yang sekiranya berkaitan dengan tujuan, rumusan, batasan, dan metode penelitian.
   * 1. Jenis dan Sumber Data

Jenis penelitian ini menggunakan data kualitatif berupa teks tertulis dari terjemahan Hadith Bukhari dalam bahasa Indonesia. Teks hadith yang akan dianalisis merupakan kumpulan teks naratif yang memiliki struktur kalimat berbeda-beda, sehingga cocok untuk pemodelan topik berbasis Natural Language Processing (NLP). Teks ini akan diproses secara kuantitatif untuk keperluan pemodelan topik.

Sumber primer data utama dalam penelitian ini adalah teks terjemahan Hadith Bukhari dalam bahasa Indonesia. Sumber sekunder dalam penelitian ini adalah Literatur dan artikel akademis tentang pemodelan topik, NLP, dan penerapan metode BerTopic, LSA, dan LDA pada teks bahasa Indonesia atau teks keagamaan. Sumber-sumber ini akan digunakan untuk memperkuat kajian pustaka dan metode analisis.

Sumber data ini dipilih untuk memastikan teks Hadith Bukhari yang digunakan akurat dan sesuai dengan standar penerjemahan yang diakui secara resmi.

* + 1. Perancangan Penelitian



**Gambar 3.1.** Perancangan Penelitian

Identifikasi masalah dan tujuan penelitian ➔

Pengumpulan dan persiapan data teks Hadith Bukhari ➔

Preprocessing teks ➔

Pemodelan topik (BerTopic, LSA, LDA) ➔

Evaluasi hasil model ➔

Visualisasi hasil ➔

Kesimpulan dan rekomendasi.

Berdasarkan diagram alir penelitian, langkah-langkah pada penelitian ini terdiri atas sebagai berikut:

1. Mengumpulkan Data

Pada tahap ini dilakukan pengumpulan teks terjemahan Hadith Bukhari dalam bahasa Indonesia dari sumber-sumber yang dapat diandalkan, seperti Kementerian Agama RI atau situs digital Islami.

Menyusun dataset teks hadith yang akan dianalisis, memastikan setiap hadith dibagi ke dalam dokumen individual untuk memudahkan analisis.Analisis Deskriptif.

1. *Preprocessing Data*

*Preprocessing* berguna untuk menyeleksi data dan mengubahnya menjadi data yang terstruktur. Berikut tahapan *Preprocessing Data.*

1. Melakukan pembersihan data dengan menghapus karakter khusus, tanda baca, dan elemen-elemen yang tidak relevan.
2. Tokenisasi teks dengan memecah teks menjadi unit kata-kata yang lebih kecil.
3. Normalisasi kata dengan mereduksi kata menjadi bentuk dasar melalui proses stemming atau lemmatization.
4. Penghapusan stopwords: menghapus kata-kata umum yang tidak memberikan informasi penting dalam analisis, seperti "di", "dan", "ke", dan sebagainya.
5. Menyiapkan teks untuk dianalisis dengan masing-masing model pemodelan topik.
6. Tahap Pemodelan Topik

Pada tahap ini, teks yang sudah diproses akan dianalisis menggunakan tiga metode pemodelan topik, BerTopic, LSA, dan LDA. Setiap model akan diimplementasikan untuk menghasilkan topik-topik berdasarkan frekuensi kemunculan kata dan distribusi tema dalam teks.

Langkah-langkah yang dilakukan adalah:

1. BerTopic

Melatih model dengan pendekatan transformer untuk menghasilkan topik-topik yang kohesif dan relevan dari teks hadith.

1. LSA

Melakukan dekomposisi matriks untuk mengidentifikasi hubungan semantik antar kata-kata dan menghasilkan topik.

1. LDA

Menerapkan model probabilistik untuk menemukan distribusi topik berdasarkan pola kata dalam dokumen hadith.

1. Evaluasi dan Hasil

Setelah pemodelan dilakukan, hasil dari ketiga model akan dievaluasi untuk menilai performanya. Metrik evaluasi utama yang digunakan dalam perbandingan ini adalah koherensi topik (*topic coherence*), yang mengukur seberapa baik topik yang dihasilkan oleh setiap model.

1. Visualisasi Hasil

Hasil dari setiap model akan divisualisasikan dalam bentuk grafik topik atau diagram hubungan antar topik untuk memudahkan pemahaman dan interpretasi. Visualisasi ini penting untuk melihat pola distribusi topik secara jelas.

Langkah-langkah:

1. Membuat visualisasi menggunakan tools seperti pyLDAvis untuk model LDA, atau library visualisasi yang mendukung BerTopic dan LSA.
2. Menyajikan topik dalam bentuk diagram distribusi atau word clouds.
3. Penutup (Kesimpulan dan Saran)

Tahap akhir penelitian adalah menyusun kesimpulan berdasarkan hasil evaluasi dan perbandingan ketiga model. Berdasarkan temuan, rekomendasi tentang metode terbaik untuk pemodelan topik Hadith Bukhari akan diberikan.

* + 1. Metode Analisis Data
  1. Preprocessing Data Teks

Dalam penelitian ini, Tahap pertama dalam analisis data adalah preprocessing untuk mempersiapkan teks Hadith Bukhari terjemahan bahasa Indonesia agar siap untuk pemodelan topik. Preprocessing melibatkan beberapa langkah berikut:

1. Tokenisasi: Memecah teks menjadi unit kata atau frasa.
2. Stopwords Removal: Menghapus kata-kata umum dalam bahasa Indonesia yang tidak membawa informasi penting (seperti "dan", "di", "yang").
3. Lemmatization: Mengubah kata-kata menjadi bentuk dasarnya (contoh: “berlari” menjadi “lari”).
4. Normalisasi Teks: Membersihkan teks dari tanda baca, angka, dan karakter khusus lainnya yang tidak diperlukan dalam pemodelan.
   1. Pemodelan Topik denganBerTopic, LSA, dan LDA

Tahap selanjutnya Pada tahap ini, teks yang telah diproses akan dimasukkan ke dalam tiga model pemodelan topik, yaitu BerTopic, LSA, dan LDA. Setiap model akan digunakan untuk menemukan topik yang terdapat dalam teks terjemahan Hadith Bukhari.

1. BerTopic: Model ini menggunakan transformer untuk memahami konteks semantik yang lebih mendalam. Setelah teks diproses, model ini akan digunakan untuk menghasilkan topik dan subtopik. Hasilnya berupa distribusi topik dan daftar kata kunci yang terkait dengan setiap topik.
2. Latent Semantic Analysis (LSA): Menggunakan dekomposisi matriks untuk menemukan pola semantik dalam hubungan antar kata dan menghasilkan topik berdasarkan representasi vektor kata.
3. Latent Dirichlet Allocation (LDA): Model ini bekerja dengan mengasumsikan bahwa setiap dokumen memiliki distribusi topik dan setiap topik memiliki distribusi kata. LDA menggunakan pendekatan probabilistik untuk menemukan distribusi tersebut dan menghasilkan topik.
   * 1. Evaluasi Model dengan Metrik Koherensi Topik

Untuk menilai kinerja masing-masing model, data yang dihasilkan akan dievaluasi menggunakan metrik koherensi topik. Koherensi topik mengukur seberapa baik kata-kata dalam topik yang dihasilkan oleh model berkaitan satu sama lain secara semantik.

Langkah-langkah evaluasi koherensi:

1. Menghitung koherensi intratopik: Mengukur keterkaitan antar kata dalam topik yang sama. Semakin tinggi koherensinya, semakin baik topik tersebut dalam merepresentasikan makna.
2. Evaluasi koherensi manual: Pakar atau akademisi di bidang hadith dapat dilibatkan untuk menilai relevansi topik yang dihasilkan dengan konteks tematik hadith.
3. Metrik koherensi yang umum digunakan adalah UMass coherence atau C\_v coherence, yang masing-masing memiliki formula spesifik untuk menghitung keterkaitan antar kata dalam topik.
   * 1. Perbandingan Hasil Model

Setelah evaluasi dilakukan, langkah selanjutnya adalah membandingkan hasil dari ketiga model (BerTopic, LSA, dan LDA) berdasarkan:

1. Koherensi Topik: Seberapa baik topik yang dihasilkan berkaitan secara semantik dan kohesif.
2. Jumlah Topik yang Dihasilkan: Membandingkan jumlah topik optimal yang dihasilkan oleh masing-masing model.
3. Interpretasi Topik: Menilai seberapa mudah topik yang dihasilkan dapat diinterpretasikan oleh manusia, terutama dalam konteks teks keagamaan seperti Hadith Bukhari.
4. Relevansi Tematik: Menilai relevansi topik yang dihasilkan dengan tema keagamaan dalam Hadith Bukhari.
   * 1. Visualisasi Hasil Topik

Untuk mempermudah interpretasi, hasil pemodelan topik dari setiap model akan divisualisasikan. Visualisasi ini bertujuan untuk mempresentasikan distribusi topik dan hubungan antar kata dalam topik dengan cara yang lebih intuitif.

Langkah-langkah visualisasi:

1. Word Clouds: Menampilkan kata-kata kunci utama dari setiap topik yang dihasilkan oleh model.
2. Topic Mapping: Menggunakan diagram seperti pyLDAvis (untuk LDA) untuk memperlihatkan hubungan antar topik dan distribusi topik dalam korpus.
3. Hierarchical Topic Tree: Visualisasi yang menunjukkan struktur subtopik dalam model BerTopic, jika dihasilkan.
   * 1. InterpretasidanAnalisis Manual

Hasil yang dihasilkan oleh model akan dianalisis secara manual oleh peneliti dan pakar di bidang hadith untuk memastikan bahwa topik yang ditemukan benar-benar relevan dengan konteks Hadith Bukhari. Proses ini melibatkan:

1. Melihat kesesuaian antara topik yang ditemukan dengan tema-tema utama dalam Hadith Bukhari (misalnya tema hukum, keluarga, pendidikan, ibadah).
2. Mengidentifikasi kata-kata kunci dalam topik dan menilai apakah kata-kata tersebut memang menggambarkan inti tema yang ditemukan*.*
   * 1. Penyusunan Kesimpulan

Tahap akhir analisis data adalah menarik kesimpulan berdasarkan perbandingan hasil dari ketiga model. Kesimpulan akan meliputi:

1. Model mana yang paling efektif dalam mengidentifikasi topik pada terjemahan Hadith Bukhari.
2. Kelebihan dan kekurangan dari masing-masing metode.
3. Rekomendasi untuk penggunaan model pemodelan topik pada teks keagamaan di masa mendatang.
4. Dengan metode analisis data ini, penelitian diharapkan dapat memberikan gambaran komprehensif tentang efektivitas dan relevansi model pemodelan topik dalam mengidentifikasi dan mengklasifikasikan tema pada teks Hadith Bukhari terjemahan Indonesia.

DAFTAR PUSTAKA

Afandi, I. R., & dkk. (2022). Analisis Sentimen Opini Masyarakat Terkait Pelayanan Jasa Ekspedisi Anteraja Dengan Metode Naive Bayes. *Jurnal Linguistik Komputasional (JLK), 5*(2), 63-69.

Alita, D., & Rahman, A. (2020). Pendeteksian Sarkasme pada Proses Analisis Sentimen Menggunakan Random Forest Classifier. *Jurnal Komputasi, 8*(2), 50-58.

Cuatanto, R., & Sutomo, R. (Indonesian Journal of Computer Science). Comparative Analysis Between Naïve Bayes Algorithm and Decision Tree Loss Rate from Fire Disaster Data in DKI Jakarta Province. *Indonesian Journal of Computer Science, 12*(4), 1787-1796.

Ferryawan, R., Kusrini, & Wibowo, F. W. (2019). Analisis Sentimen Wisata Jawa Tengah Menggunakan Naϊve Bayes. *Jurnal INFORMA, 5*(3), 55-60.

Flores, V. A., Permatasari, P. A., & Jasa, L. (2020). Penerapan Web Scraping Sebagai Media Pencarian dan Menyimpan Artikel Ilmiah Secara Otomatis Berdasarkan Keyword. *Majalah Ilmiah Teknologi Elektro, 19*(2), 157-162. doi:10.24843/MITE.2020.v19i02.P06

Giovani, A. P., Ardiansyah, Haryanti, T., Kurniawati, L., & Gata, W. (2020). ANALISIS SENTIMEN APLIKASI RUANG GURU DI TWITTER MENGGUNAKAN ALGORITMA KLASIFIKASI. *Jurnal TEKNOINFO, 14*(2), 116-124. doi:10.33365/jti.v14i2.679

Harahap, F., & dkk. (2021). Penerapan Data Mining Dengan Algoritma Naive Bayes Classifier Dalam Memprediksi Pembelian Cat. *Jurnal Ilmiah Informatika, 9*(1).

Harun, A., & Ananda, D. P. (2021). Analisa Sentimen Opini Publik Tentang Vaksinasi Covid-19 di Indonesia Menggunakan Naïve Bayes dan Decision Tree. *MALCOM: Indonesian Journal of Machine Learning and Computer Science, 1*(1), 58-63.

Irnawati, O., & Solecha, K. (2022). Analisis Sentimen Ulasan Aplikasi Flip Menggunakan Naïve Bayes dengan Seleksi Fitur PSO. *Jurnal Ilmiah Intech, 4*(2), 189-199.

Juniarsih, S., Ripantia, E. F., & Pratama, E. E. (2020). Implementasi Naive Bayes Classifier pada Opinion Mining Berdasarkan Tweets Masyarakat Terkait Kinerja Presiden dalam Aspek Ekonomi. *JUSTIN (Jurnal Sistem dan Teknologi Informasi), 8*(3), 239-246. doi:10.26418/justin.v8i3.39118

Lestari, N. A., Akhriza, T. M., & Yuniar, E. (2020). METODE NAÏVE BAYES CLASSIFIER DENGAN TEXTBLOB UNTUK ANALISIS SENTIMEN TERHADAP PELAYANAN INDIHOME DAN FIRST MEDIA. *Seminar Nasional Teknologi Informasi dan Komunikasi STI&K (SeNTIK), 4*(1), 283-290.

Nishfi Ilmiah Huda, D., Prianto, C., & Awangga, R. M. (2023). Analisis Sentimen Perbandingan Layanan Jasa Pengiriman Kurir Pada Ulasan Play Store Menggunakan Metode Random Forest dan Decision Tree. *Jurnal Ilmiah Informatika, 11*(2), 151-155.

Nitami, M. T., & Februariyanti, H. (2022). Analisis Sentimen Ulasan Ekspedisi J&T Express Menggunakan Algoritma Naive Bayes. *Jurnal Manajemen Informatika & Sistem Informasi (MISI), 5*(1), 20-29. doi:10.36595/misi.v5i1

Pertiwi, M. W. (2019). Analisis Sentimen Opini Publik Mengenai Sarana dan Transportasi Mudik Tahun 2019 pada Twitter Menggunakan Algoritma Naïve Bayes, Neural Network, KNN dan SVM. *Inti Nusa Mandiri, 14*(1), 27-32.

Putri, D. S., & Ridwan, T. (2023). Analisis Sentimen Ulasan Aplikasi Pospay dengan Algoritma Support Vector Machine. *Jurnal Ilmiah Informatika, 11*(1), 33-40.

Ramadhani, T., Sari, Y. A., & Santoso, E. (2021). Analisis Sentimen Masyarakat Indonesia terhadap Covid-19 pada Media Sosial Twitter menggunakan Metode Naïve Bayes. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, 5*(12), 5680-5686.

Ratnawati, F. (2018). Implementasi Algoritma Naive Bayes Terhadap Analisis Sentimen Opini Film Pada Twitter. *JURNAL INOVTEK POLBENG - SERI INFORMATIKA, 3*(1), 50-55.

Romadloni, N. T., Santoso, I., & Budilaksono, S. (2019). Perbandingan Metode Naive Bayes, KNN dan Decision Tree Terhadap Analisis Sentimen Transportasi KRL Commuter Line. *Jurnal IKRA-ITH Informatika, 3*(2), 1-8.

Romadoni, F., Umaidah, Y., & Sari, B. N. (2020). Text Mining Untuk Analisis Sentimen Pelanggan Terhadap Layanan Uang Elektronik Menggunakan Algoritma Support Vector Machine. *Jurnal SISFOKOM (Sistem Informasi dan Komputer), 9*(2), 247-253. doi:10.32736/sisfokom.v9i2.903

Rosdiana, & dkk. (2019). Analisis Sentimen pada Twitter terhadap Pelayanan Pemerintah Kota Makassar. *Proceeding SNTEI*, 87-93.

Samsir, & dkk. (2021). Analisis Sentimen Pembelajaran Daring Pada Twitter di Masa Pandemi COVID-19 Menggunakan Metode Naïve Bayes. *JURNAL MEDIA INFORMATIKA BUDIDARMA, 5*(1), 157-163. doi:10.30865/mib.v5i1.2604

Saraswati, M., & Riminarsih, D. (2020). Analisis Sentimen Terhadap Pelayanan KRL Commuterline Berdasarkan Data Twitter Menggunakan Algoritma Bernoulli Naive Bayes. *Jurnal Ilmiah Komputer dan Informatika (KOMPUTA), 25*(3), 225-230. doi:https://doi.org/10.35760/ik.2020.v25i3.3256

Sari, F. V., & Wibowo, A. (2019). Analisis Sentimen Pelanggan Toko Online JD.ID Menggunakan Metode Naïve Bayes Classifier Berbasis Konversi Ikon Emosi. *Jurnal SIMETRIS, 10*(2), 681-686.

Septian, J. A., Fahrudin, T. M., & Nugroho, A. (2020). Analisis Sentimen Pengguna Twitter Terhadap Polemik Persepakbolaan Indonesia Menggunakan Pembobotan TF-IDF dan K-Nearest Neighbor. *Journal of Intelligent Systems and Computation*, 43-49.

Septiana, R. D., Susanto, A. B., & Tukiyat, T. (2021). Analisis Sentimen Vaksinasi Covid-19 Pada Twitter Menggunakan Naive Bayes Classifier Dengan Feature Selection Chi-Squared Statistic Dan Particle Swarm Optimization. *Jurnal Sistem Komputer dan Kecerdasan Buatan, 5*(1), 49-56.

Setiawan, H., Utami, E., & Sudarmawan, S. (2021). Analisis Sentimen Twitter Kuliah Online Pasca Covid-19 Menggunakan Algoritma Support Vector Machine dan Naive Bayes. *Jurnal Komtika (Komputasi dan Informatika), 5*(1), 43-51. doi:10.31603/komtika.v5i1.5189

Siti, & Mayary, J. (2020). Analisis Sentimen Pengguna Twitter terhadap Dompet Elektronik dengan Metode Lexicon Based dan K-Nearest Neighbor. *Jurnal Ilmiah Informatika Komputer, 25*(1), 1-12. doi:10.35760/ik.2020.v25i1.2411

Wahyudi, R., & Kusumawardhana, G. (2021). Analisis Sentimen pada review Aplikasi Grab di Google Play Store Menggunakan Support Vector Machine. *JURNAL INFORMATIKA, 8*(2), 200-207.

Wijaya, H. D., & Dwiasnati, S. (2020). Implementasi Data Mining dengan Algoritma Naïve Bayes pada Penjualan Obat. *JURNAL INFORMATIKA, 7*(1), 1-7.

Yasid, M., & Junaedi, L. (2019). Analisis Sentimen Maskapai Citilink Pada Twitter Dengan Metode Naïve Bayes. *JURNAL ILMIAH INFORMATIKA, 7*(2), 83-91.